

# **Modeling missing-data processes: An IRTree-based approach.**

Dries Debeer (University of Leuven)  
Rianne Janssen (University of Leuven)  
Paul De Boeck (Ohio State University)

# **This talk?**

Modeling missing item responses ..

.. in low-stake assessments ..  
(educational measurement)

.. using IRTrees.

(De Boeck & Partchev, 2012)

# Content

1. Missing data
2. Handling missing data
3. IRTree
4. Two IRTrees for missing data
5. Illustration
6. Discussion

# Content

1. Missing data
2. Handling missing data
3. IRTree
4. Two IRTrees for missing data
5. Illustration
6. Discussion

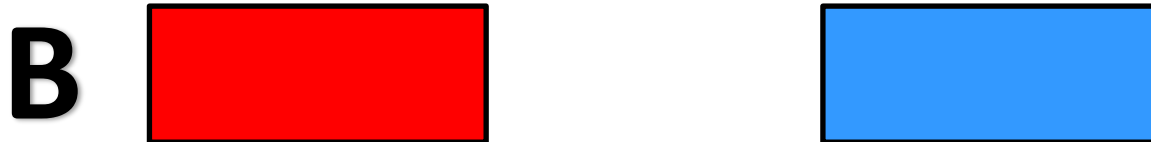
# Missing data

## Different types

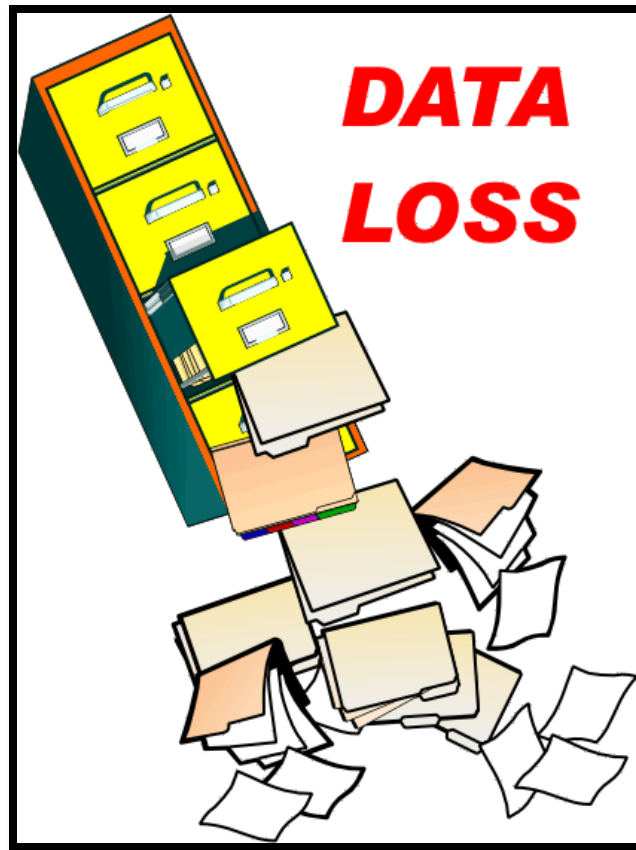
1. Before administration
2. After administration
3. During administration

# 1. Before administration

By Design



## 2. After administration



# Missing data

## Different types

1. Before administration
2. After administration
3. During administration
  - Related to test takers
  - Related to items



# 3. During administration

Two types of non-responses:

- Near the end of the assessment
  - “r”
  - “not reached”
- Well before the end of the assessment
  - “d”
  - “skipped”

# Data Matrix: $X$ ( $P \times I$ )

*Items*

*Persons*

	<i>1</i>	<i>2</i>	<i>...</i>	<i>i</i>	<i>...</i>	<i>I-3</i>	<i>I-2</i>	<i>I-1</i>	<i>I</i>
<i>1</i>	1	1	...	9	...	0	9	9	9
<i>2</i>	9	0	...	1	...	1	9	0	1
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
<i>p</i>	0	1	...	9	...	9	1	0	9
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
<i>P-1</i>	0	9	...	1	...	1	0	9	9
<i>P</i>	1	0	...	0	...	9	9	9	9

# Data Matrix: $X$ ( $P \times I$ )

*Items*

	<i>1</i>	<i>2</i>	...	<i>i</i>	...	<i>I-3</i>	<i>I-2</i>	<i>I-1</i>	<i>I</i>
<i>1</i>	1	1	...	9	...	0	9	9	9
<i>2</i>	9	0	...	1	...	1	9	0	1
⋮	⋮	⋮		⋮		⋮	⋮	⋮	⋮
<i>p</i>	0	1	...	9	...	9	1	0	9
⋮	⋮	⋮		⋮		⋮	⋮	⋮	⋮
<i>P-1</i>	0	9	...	1	...	1	0	9	9
<i>P</i>	1	0	...	0	...	9	9	9	9

*Persons*

# Data Matrix: $X$

	1	2	...	$i$	...	$I-3$	$I-2$	$I-1$	$I$
1	1	1	...	9	...	0	9	9	9
2	9	0	...	1	...	1	9	0	1
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
$p$	0	1	...	9	...	9	1	0	9
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
$p-1$	0	9	...	1	...	1	0	9	9
$p$	1	0	...	0	...	9	9	9	9

Missing responses: "9"

# Data Matrix: $X$

	1	2	...	$i$	...	$l-3$	$l-2$	$l-1$	$l$
1	1	1	...	<b>d</b>	...	0	9	9	9
2	<b>d</b>	0	...	1	...	1	<b>d</b>	0	1
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
$p$	0	1	...	<b>d</b>	...	<b>d</b>	1	0	9
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
$p-1$	0	<b>d</b>	...	1	...	1	0	9	9
$p$	1	0	...	0	...	9	9	9	9

Skipped items: "d"

# Data Matrix: $X$

	$1$	$2$	$\dots$	$i$	$\dots$	$l-3$	$l-2$	$l-1$	$l$
$1$	1	1	$\dots$	d	$\dots$	0	r	r	r
$2$	d	0	$\dots$	1	$\dots$	1	d	0	1
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
$p$	0	1	$\dots$	d	$\dots$	d	1	0	r
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
$p-1$	0	d	$\dots$	1	$\dots$	1	0	r	r
$p$	1	0	$\dots$	0	$\dots$	r	r	r	r

Not-reached items: "r"

# Data Matrix: $X$

	1	2	...	$i$	...	$l-3$	$l-2$	$l-1$	$l$
1	1	1	...	d	...	0	r	r	r
2	d	0	...	1	...	1	d	0	1
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
$p$	0	1	...	d	...	d	1	0	r
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
$p-1$	0	d	...	1	...	1	0	r	r
$p$	1	0	...	0	...	r	r	r	r

Skipped items: "d"

"r" : Not-reached items

# Content

1. Missing data
2. Handling missing data
3. IRTree
4. Two IRTrees for missing data
5. Illustration
6. Discussion



# Handling missing data

Common strategies:

Missing = Wrong

- $r/d = 0$
- Data imputation (IAS)

Ignore missing responses

- $r/d = NA$
- Assuming MAR

Combination of both

# Handling missing data

However...

Missingness might be related to

- Lower *motivation* => higher probability of *d*?
- Lower *speed* => higher probability of *r*?
- ...

⇒ The *r/d* -responses can be informative.

Missingness might be related to

- the proficiency of the test taker.

⇒ Bias on the ability measurement (MNAR)

⇒ **Not ignorable**

# Missing data: Categorisation (Rubin, 1976)

MCAR = Missing Completely At Random

MAR = Missing At Random

MNAR = Missing Not At Random

Missingness indicator  $u_{pi}$

- Whenever  $x_{pi}$  is observed,  $u_{pi} = 1$
- Whenever  $x_{pi}$  is not observed,  $u_{pi} = 0$
- $\mathbf{U} : P \times I$  matrix

# $\mathbf{X}_{obs}$ & $\mathbf{X}_{mis}$

Based on the missingness indicator matrix  $\mathbf{U}$  the *full* data matrix falls apart in:

- observed part  $\mathbf{X}_{obs}$
- unobserved part  $\mathbf{X}_{mis}$

# MNAR

The missingness does depend on the observed and the unobserved data.

$$P(U|X_{obs}, X_{mis}) = P(U|X_{obs}, X_{mis})$$

- **Non-ignorable** for likelihood based inferences
- Possible bias

**Model missing data  
process/processes**

# Modeling missing data processes

Latent variable model (Moustaki & O'Muircheartaigh, 2000)



- Skipped (Holman & Glas, 2005)
- Not reached (Glas & Pimentel, 2008)
- Both...

⇒ IRTree (De Boeck & Partchev, 2012)

# Content

1. Missing data
2. Handling missing data
3. IRTree
4. Two IRTrees for missing data
5. Illustration
6. Discussion

# IRTree (De Boeck & Partchev, 2012)

- IRT-model
- tree structure
- sequentially interconnected sub-processes
- model a response process
  - observed variables (responses) 
  - unobserved (latent) sub-processes 

Here:

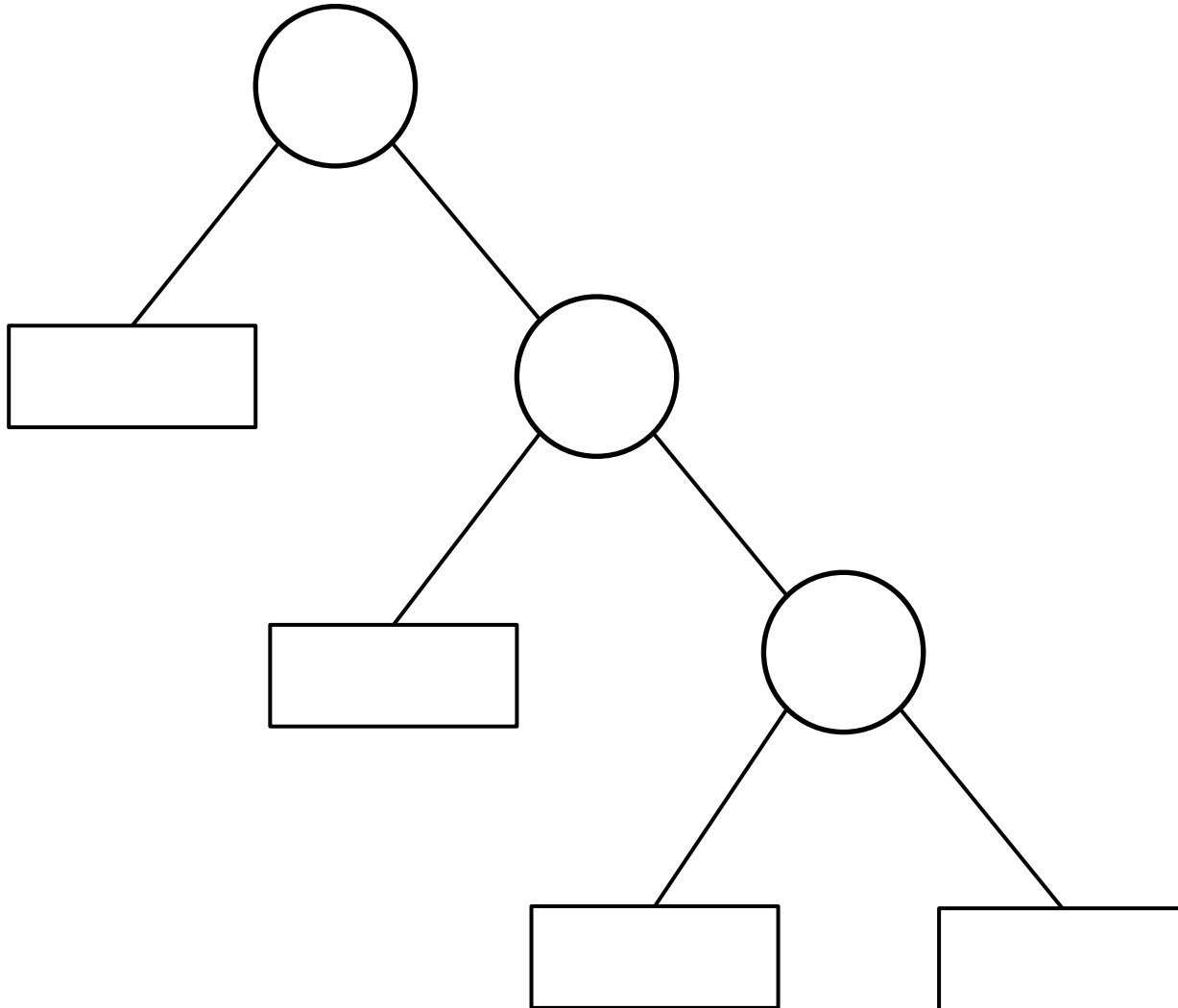
- IRTree to model response categories  $(0, 1, d, r)$
- With distinct missing data processes
- 2 IRTrees



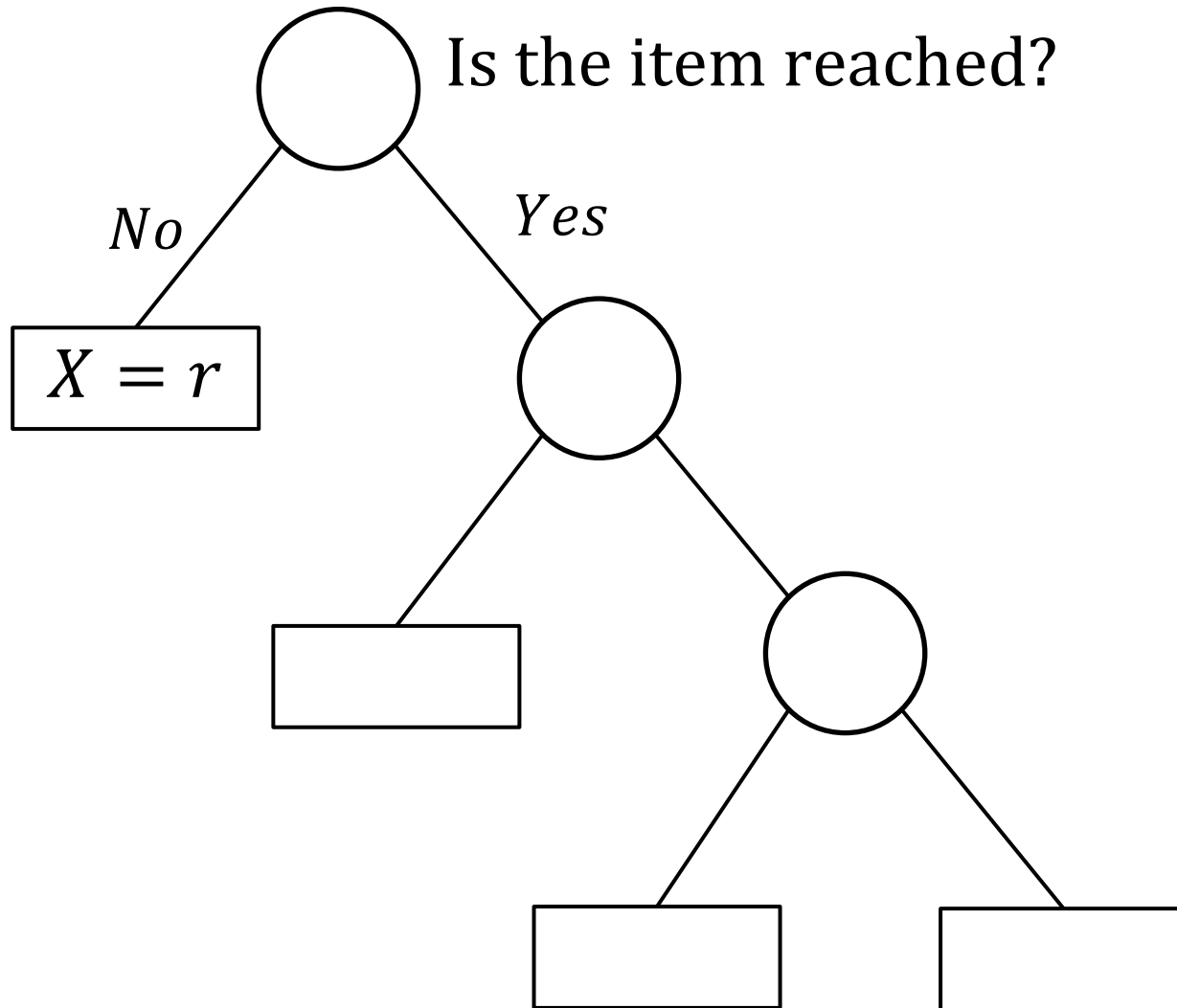
# Content

1. Missing data
2. Handling missing data
3. IRTree
4. IRTrees for missing data
5. Illustration
6. Discussion

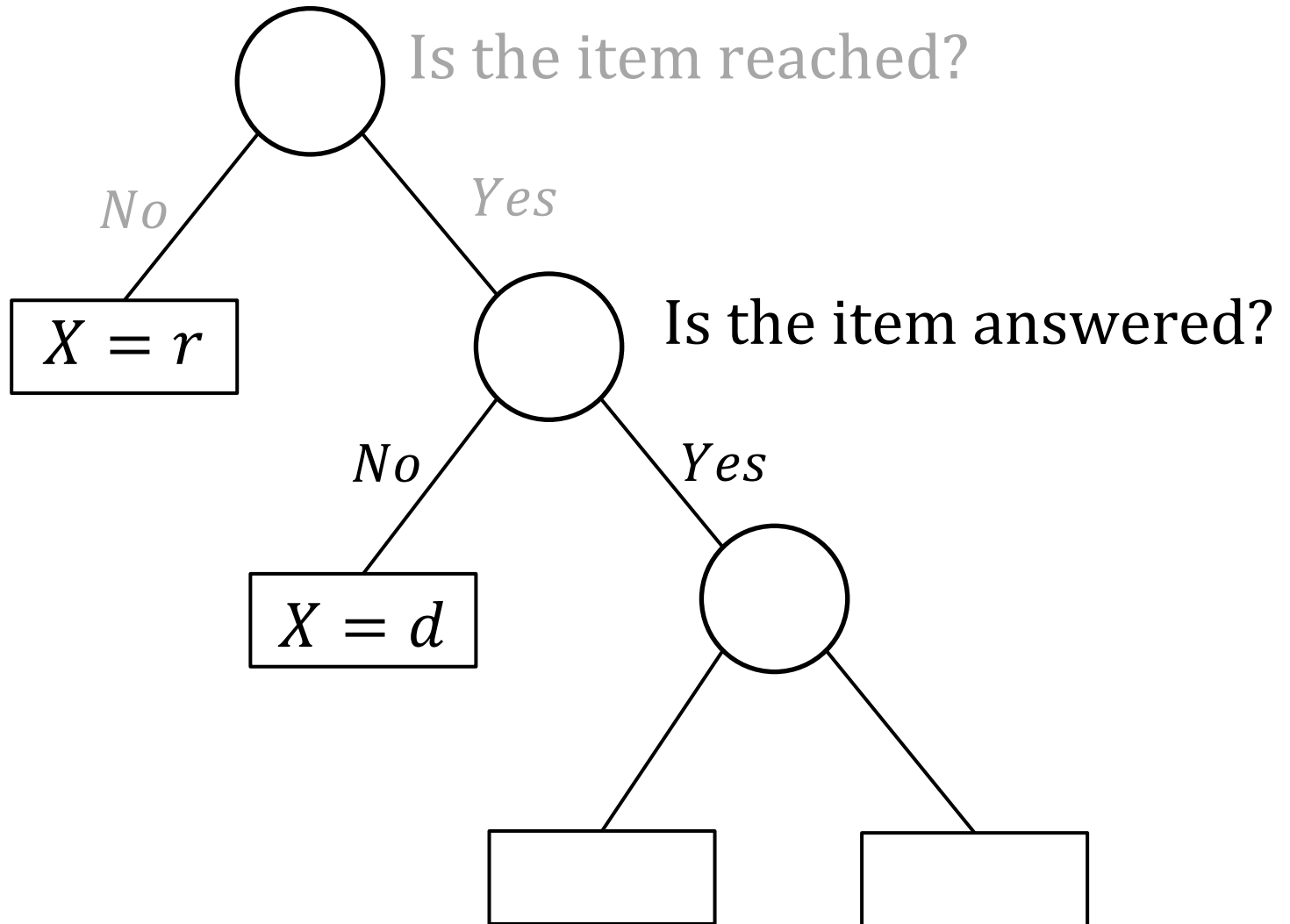
# IRTree 1



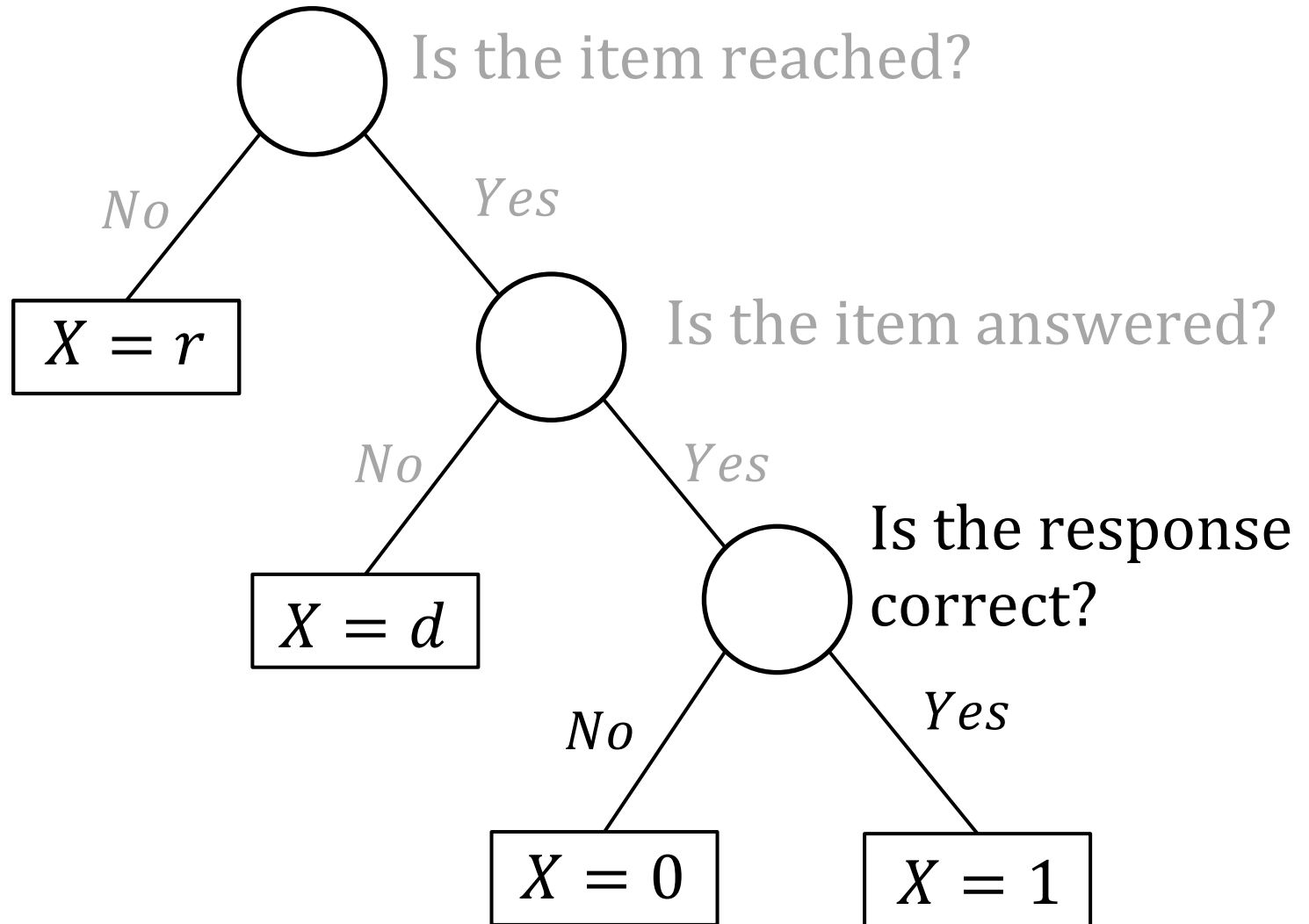
# IRTree 1



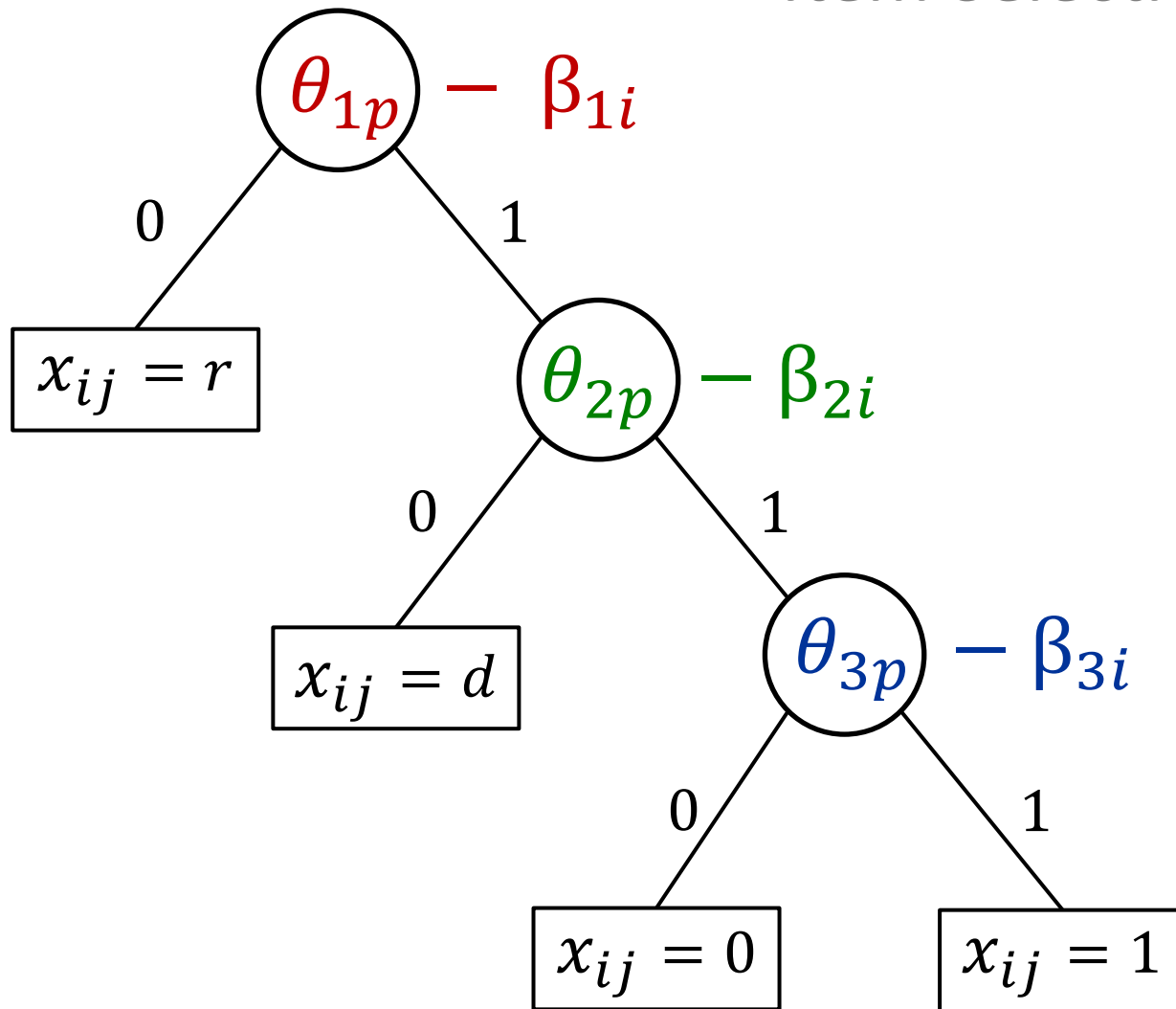
# IRTree 1



# IRTree 1 *Item Selection Model*



# Item Selection Model



# IRTree 1: Probabilities

$$P[x_{ij} = r] = 1 - \text{logit}^{-1}[\theta_{1p} - \beta_{1i}]$$

$$P[x_{ij} = d] = \text{logit}^{-1}[\theta_{1p} - \beta_{1i}] \times (1 - \text{logit}^{-1}[\theta_{2p} - \beta_{2i}])$$

$$\begin{aligned} P[x_{ij} = 0] \\ &= \text{logit}^{-1}[\theta_{1p} - \beta_{1i}] \times \text{logit}^{-1}[\theta_{2p} - \beta_{2i}] \times (1 \\ &\quad - \text{logit}^{-1}[\theta_{3p} - \beta_{3i}]) \end{aligned}$$

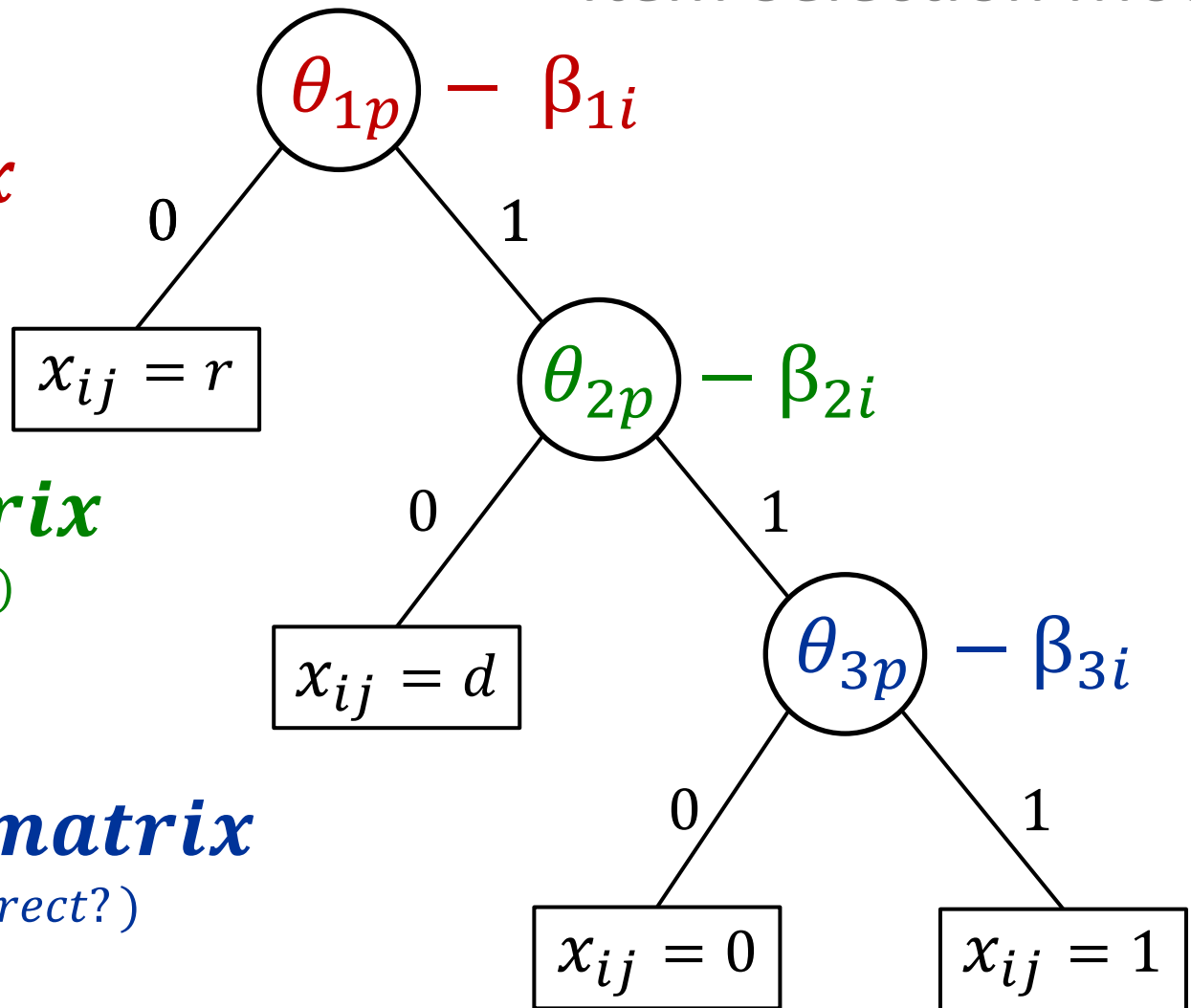
$$\begin{aligned} P[x_{ij} = 1] \\ &= \text{logit}^{-1}[\theta_{1p} - \beta_{1i}] \times \text{logit}^{-1}[\theta_{2p} - \beta_{2i}] \times \text{logit}^{-1}[\theta_{3p} \\ &\quad - \beta_{3i}] \end{aligned}$$

# Item Selection Model

**$Y_1$  – matrix**  
(reached?)

**$Y_2$  – matrix**  
(answered?)

**$Y_3$  – matrix**  
(correct?)





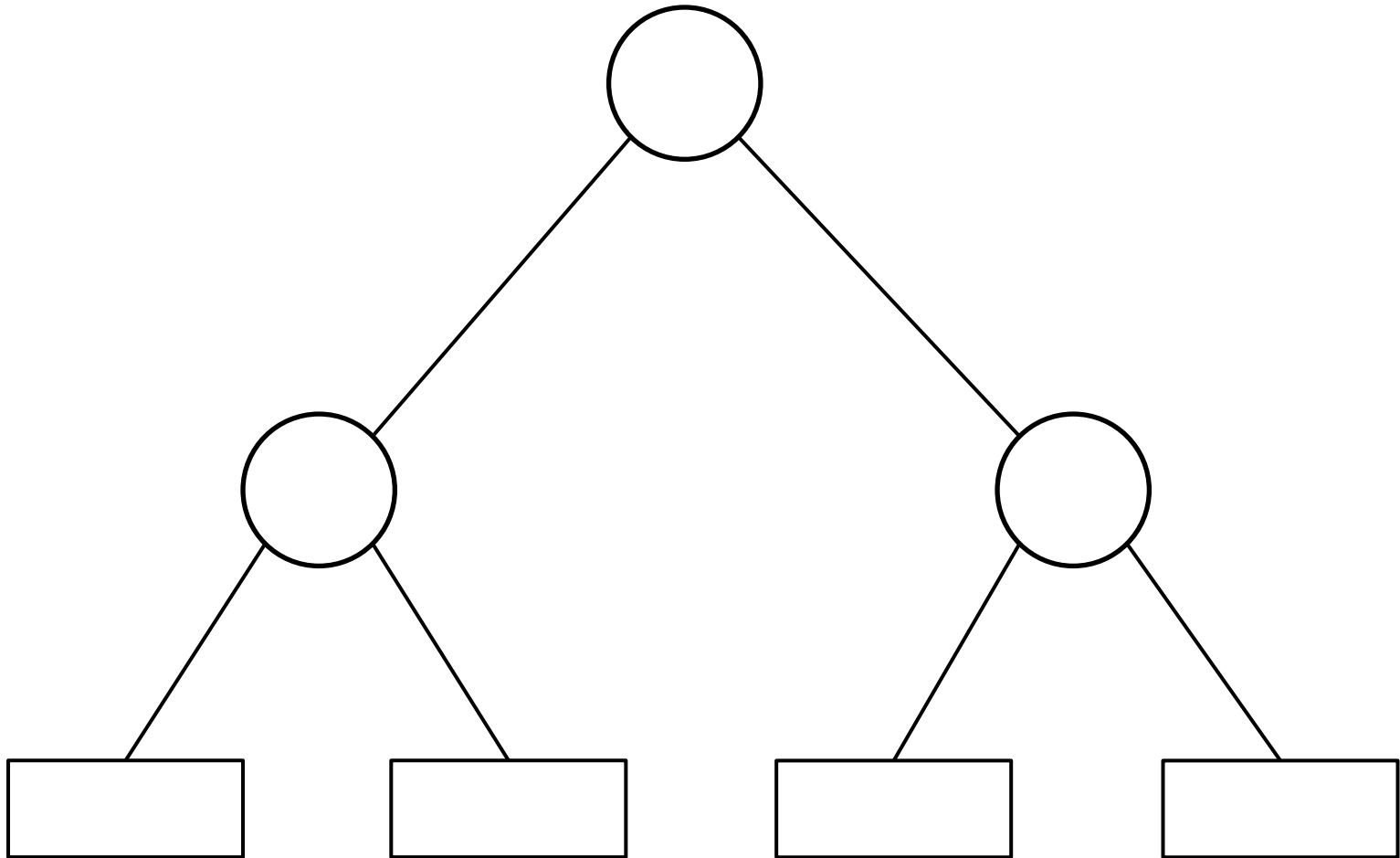
# Mapping on subitems

<i><b>Item</b></i>	<i><b>X</b></i>	<i><b>Code</b></i>	
1	1	1	
2	1	0	
3	9	d	
4	0	0	
5	9	r	
6	9	r	
7	9	r	

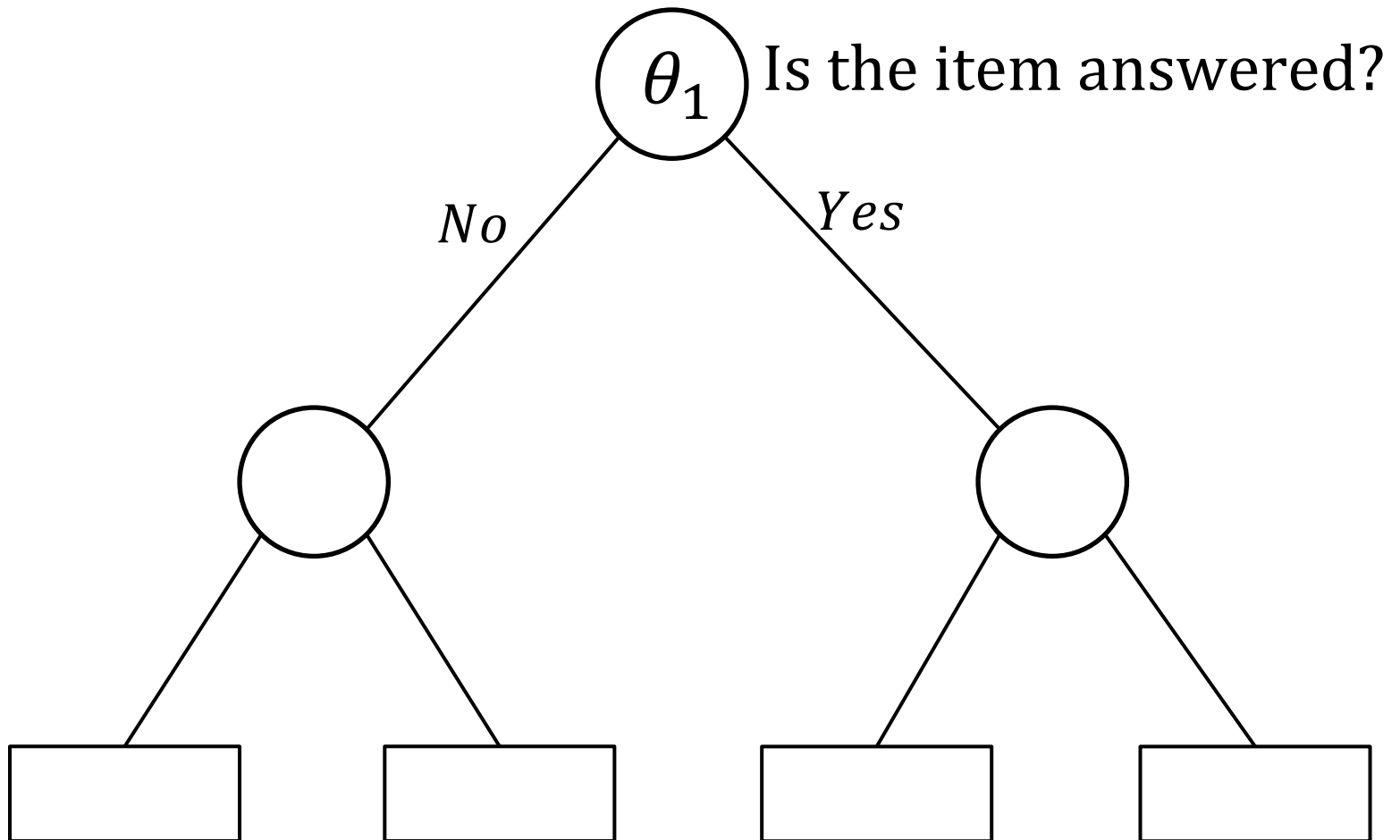
# Mapping on subitems

<i>Item</i>	<i>X</i>	<i>Code</i>	<i>Y<sub>1</sub></i>	<i>Y<sub>2</sub></i>	<i>Y<sub>3</sub></i>
1	1	1	1	1	1
2	1	0	1	1	1
3	9	d	1	0	NA
4	0	0	1	1	0
5	9	r	0	NA	NA
6	9	r	NA	NA	NA
7	9	r	NA	NA	NA

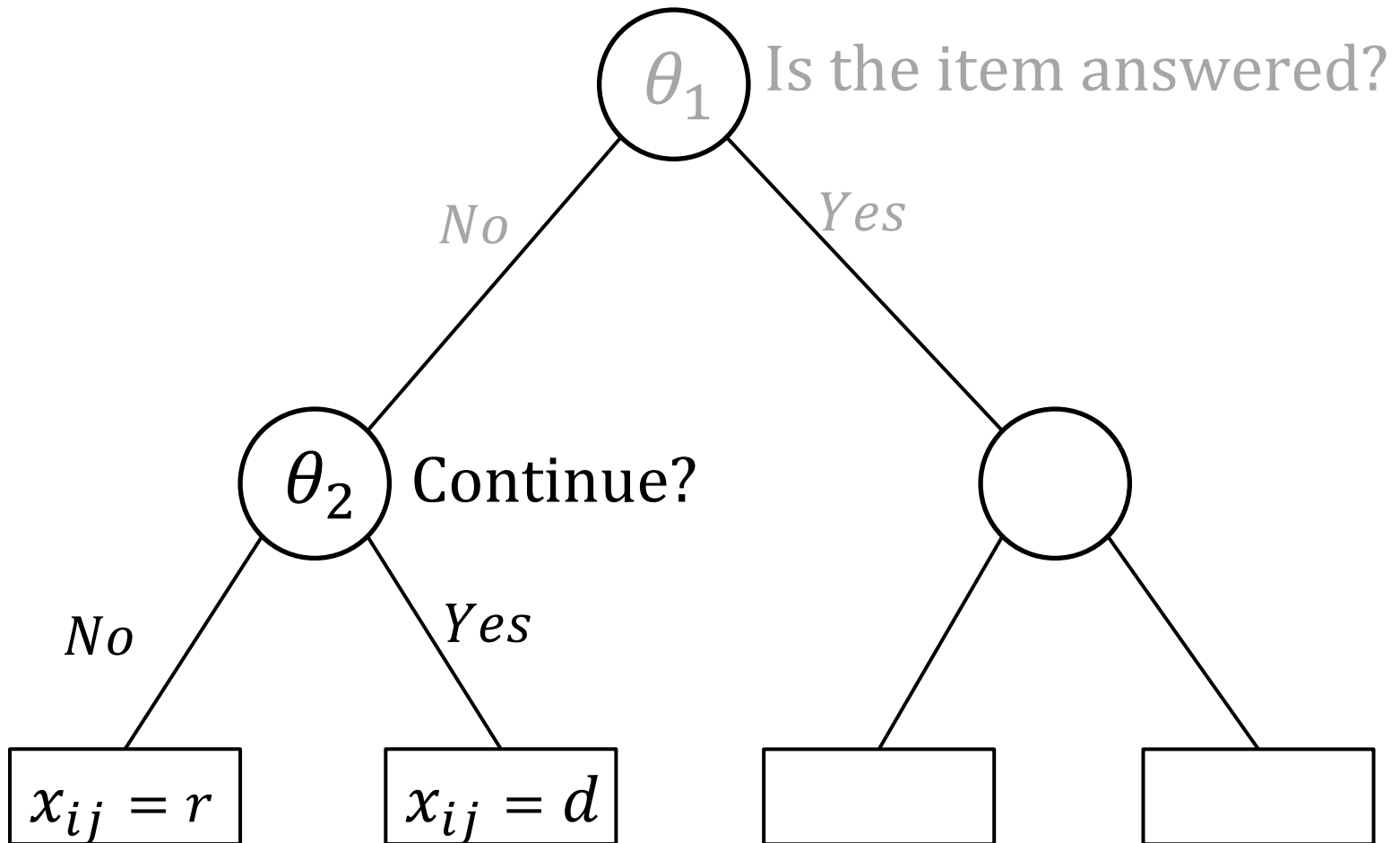
# IRTree 2



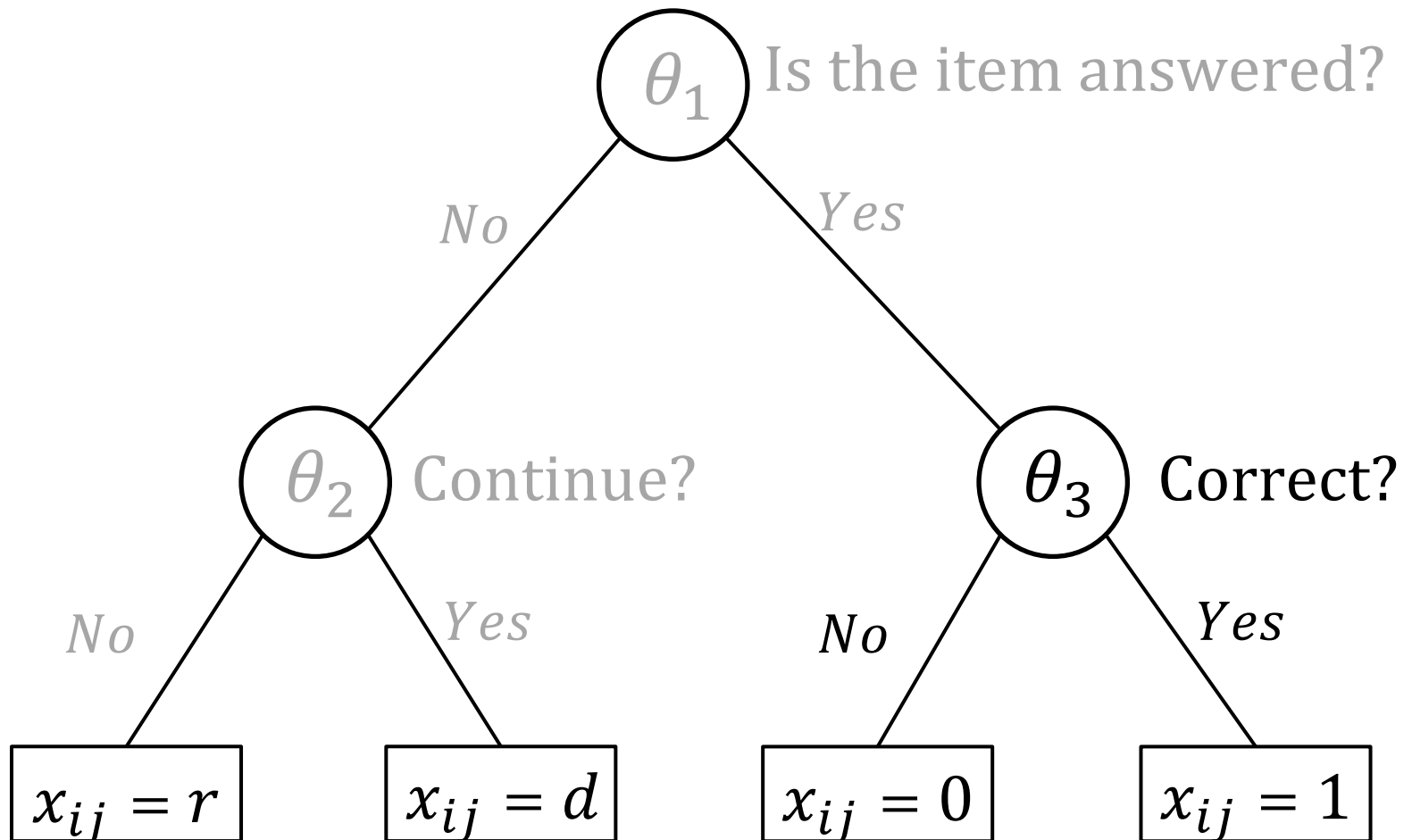
# IRTree 2



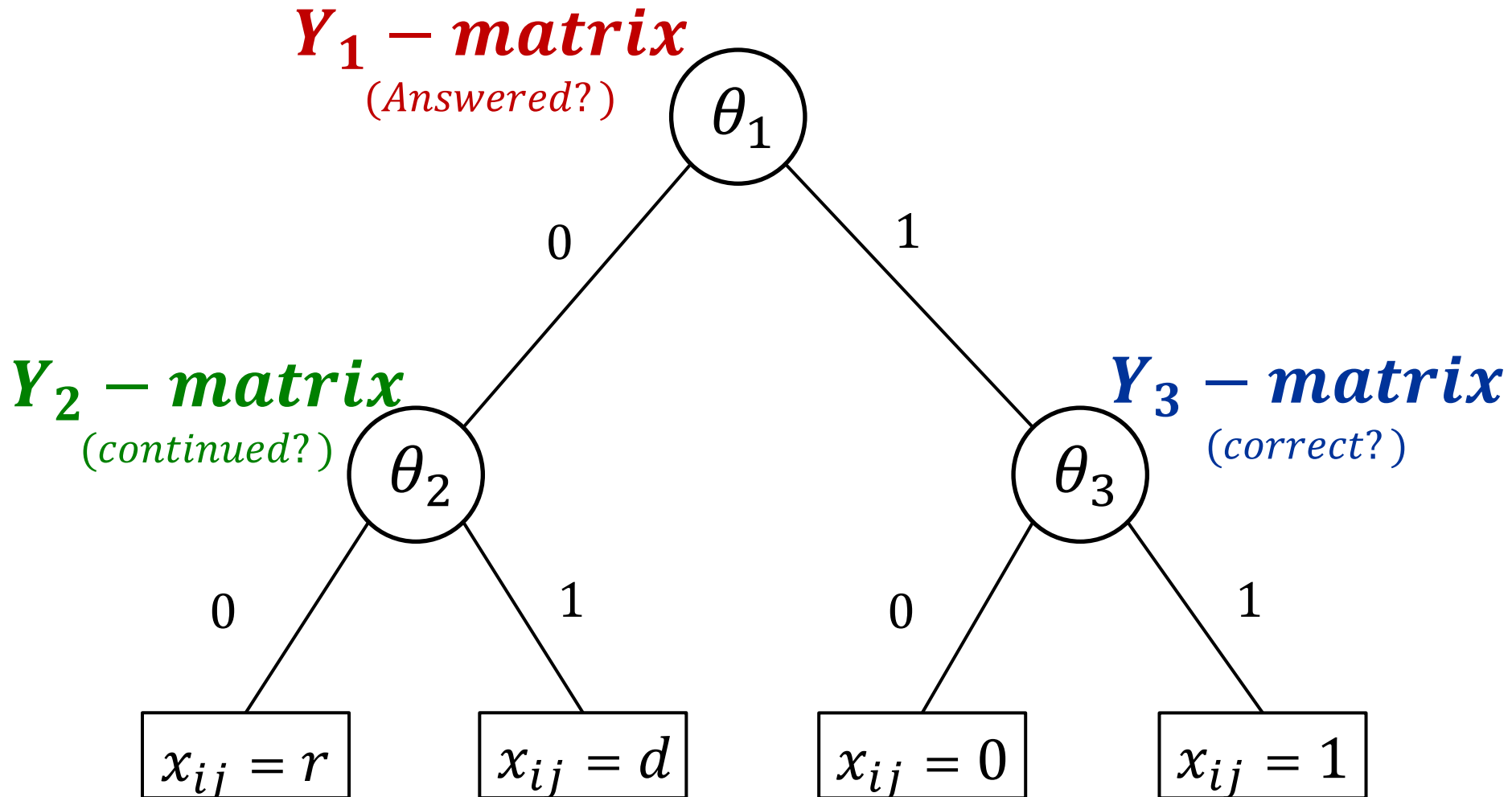
# IRTree 2



## IRTree 2 *Continuing Effort Model*



# Continuing Effort Model



## IRTree 2: Probabilities

$$P[x_{ij} = r] = (1 - \text{logit}^{-1}[\theta_{1p} - \beta_{1i}]) \times (1 - \text{logit}^{-1}[\theta_{2p} - \beta_{2i}])$$

$$P[x_{ij} = d] = (1 - \text{logit}^{-1}[\theta_{1p} - \beta_{1i}]) \times \text{logit}^{-1}[\theta_{2p} - \beta_{2i}]$$

$$P[x_{ij} = 0] = \text{logit}^{-1}[\theta_{1p} - \beta_{1i}] \times (1 - \text{logit}^{-1}[\theta_{3p} - \beta_{3i}])$$

$$P[x_{ij} = 1] = \text{logit}^{-1}[\theta_{1p} - \beta_{1i}] \times \text{logit}^{-1}[\theta_{3p} - \beta_{3i}]$$



# Mapping on subitems

<i>Item</i>	<i>X</i>	<i>Code</i>	<i>Y<sub>1</sub></i>	<i>Y<sub>2</sub></i>	<i>Y<sub>3</sub></i>
1	1	1	1	NA	1
2	1	0	1	NA	1
3	9	d	0	1	NA
4	0	0	1	NA	0
5	9	r	0	0	NA
6	9	r	NA	NA	NA
7	9	r	NA	NA	NA

# 1PL IRTree

$$\pi(x_{pi} = m | \boldsymbol{\theta}_p, \boldsymbol{\beta}_i) = \prod_{r=1}^R \left[ \frac{\exp(\theta_p^{(n)} + \beta_i^{(n)})^{t_{mn}}}{1 + \exp(\theta_p^{(n)} + \beta_i^{(n)} \beta_i^{(n)})} \right]^{d_{mn}}$$

$m = (0, 1, d, r)$  (response categories)

$n = (1, 2, 3)$  (nodes)

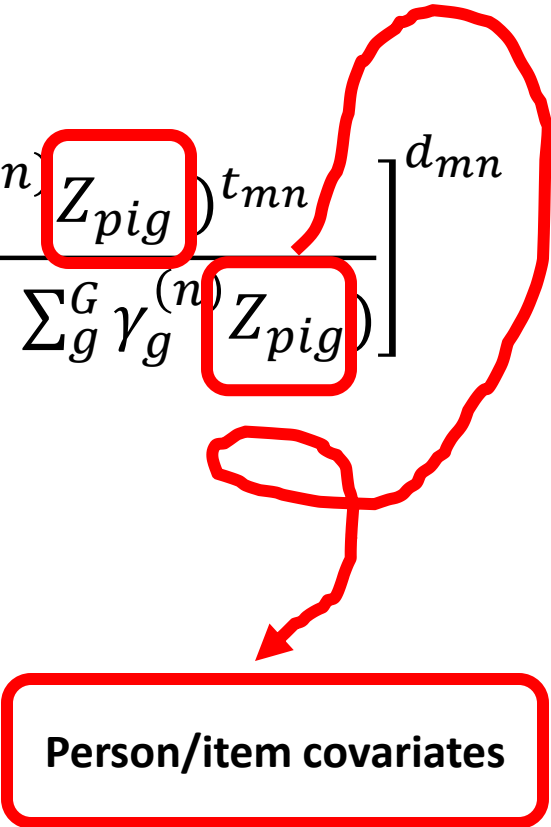
$t_{mn} = (0, 1, NA)$  mapping of  $m$  on  $n$

$d_{mn} = \begin{cases} 0 & \text{when } t_{mr} = NA \\ 1 & \text{otherwise} \end{cases}$

**Family of GLMM**

# IRTrees

$$\pi(x_{pi} = m | \boldsymbol{\theta}_p, \boldsymbol{\beta}_i)$$

$$= \prod_{r=1}^R \left[ \frac{\exp(\theta_p^{(n)} + \beta_i^{(n)} + \sum_g^G \gamma_g^{(n)} Z_{pig})^{t_{mn}}}{1 + \exp(\theta_p^{(n)} + \beta_i^{(n)} + \sum_g^G \gamma_g^{(n)} Z_{pig})} \right]^{d_{mn}}$$


$m = (0, 1, d, r)$  (response categories)

$n = (1, 2, 3)$  (nodes)

$t_{mr} = (0, 1, NA)$  mapping of  $m$  on  $n$

$d_{mr} = \begin{cases} 0 & \text{when } t_{mr} = NA \\ 1 & \text{otherwise} \end{cases}$

Person/item covariates

# IRTrees

Relation between missing data processes can be tested

- Correlation between latent traits
- Correlation between item parameters
- MNAR
- Missing = wrong
- Multinomial model

# Interpretation IRTree

IRTree1: *Item Selection Model*

- For more speeded tests?

IRTree2: *Continuing Effort Model*

- Motivation / fatigue?

Proficiency process: the same

Missing data processes: different

# Which IRTree?

- Proficiency
- Test
- Administering conditions
- Model fit (information criterion)

# Content

1. Missing data
2. Handling missing data
3. IRTree
4. Two IRTrees for missing data
5. Illustration
6. Discussion

# Illustration: Fast arithmetic

Part of national math survey in the Flemish part of Belgium

$N = 2288$  students

Speed test (easy items)



# Design

+

20 items

40 sec

-

20 items

40 sec

×

20 items

50 sec

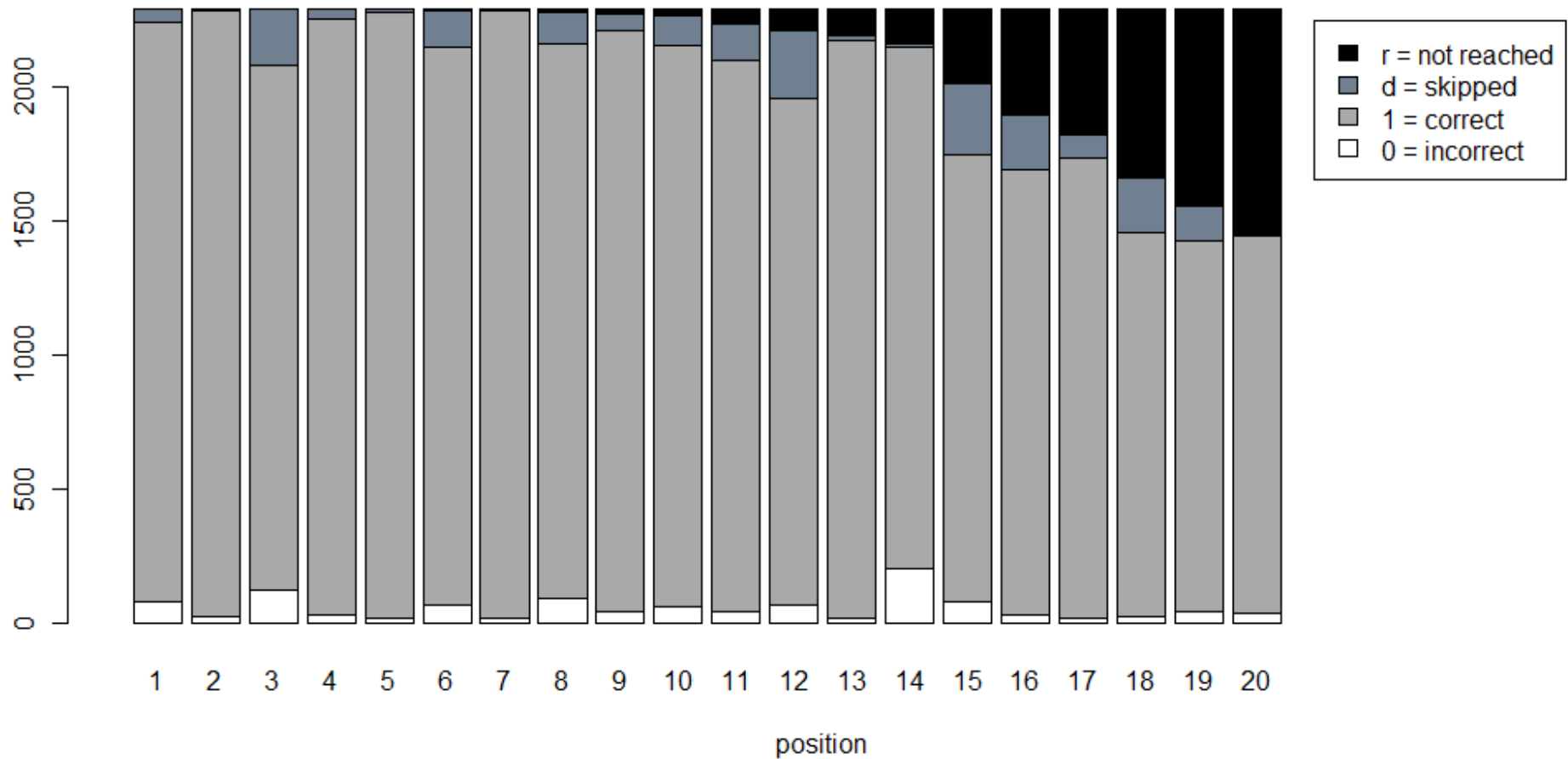
:

20 items

50 sec

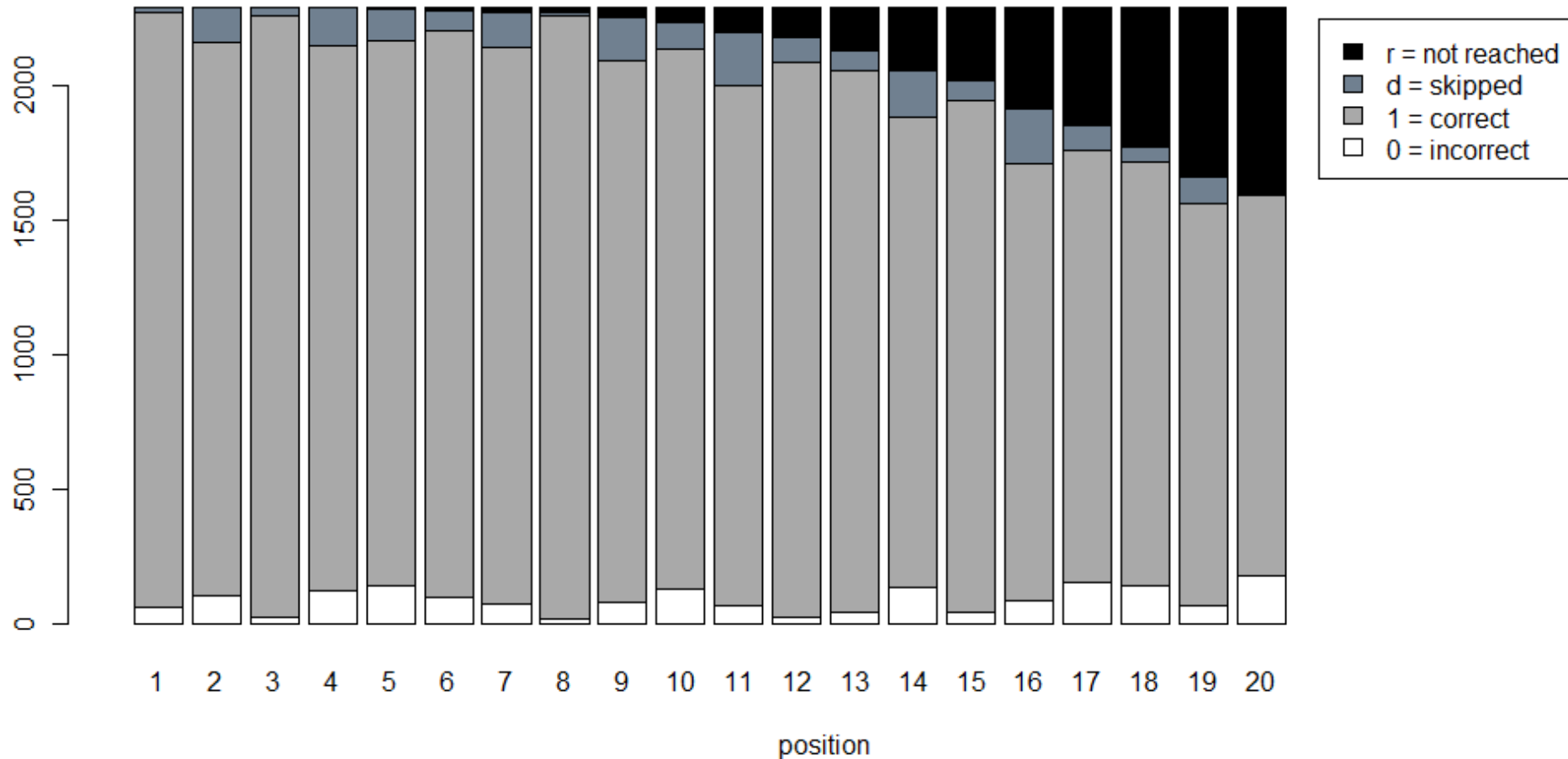
# Descriptives

Responses per item position - Multiplication



# Descriptives

Responses per item position - Division



# Estimation

In R (Lmer)

$$(\theta_p^1, \theta_p^2, \theta_p^3) \sim MNV(\mathbf{0}, \mathbf{\Sigma}_\theta)$$

$$(\varepsilon_i^1, \varepsilon_i^2, \varepsilon_i^3) \sim MNV(\mathbf{0}, \mathbf{\Sigma}_\varepsilon)$$

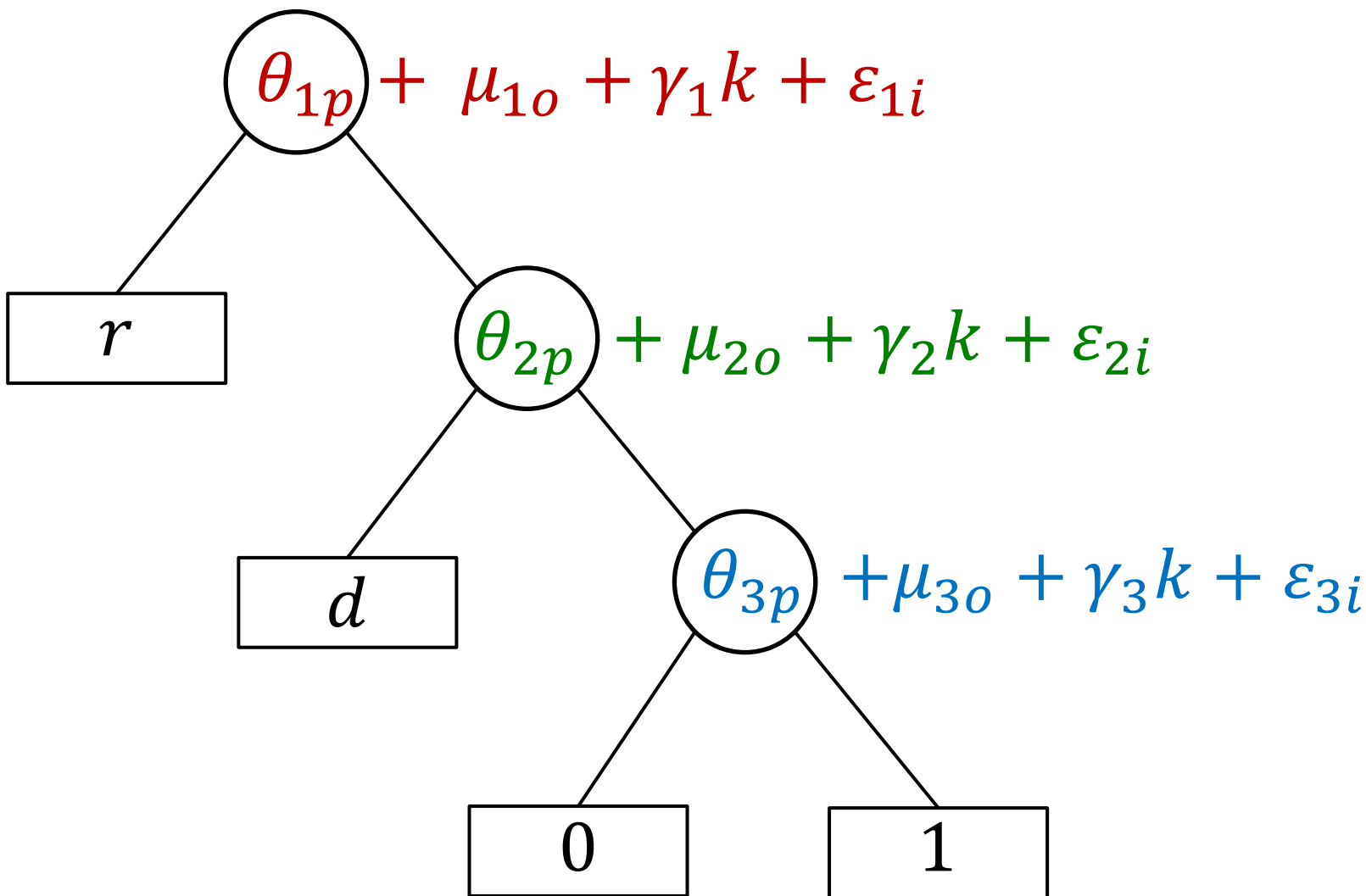
- ⇒ person / item correlations are estimated
- ⇒ relation between the processes?

# Which IRTree?

Speed tests: *Item selection model*

Model	- logL	AIC	BIC
Item Selection	44137	88314	88516
Continuing Effort	44563	89167	89369

# Final model



# Results

Node ( $n$ )	Position effect			Change in odds	
	$\gamma^{(n)}$	$SE(\gamma^{(n)})$	$p$	+1 position	+5 positions
Dropping out (1)	-0.461	.0143	< .001	1.59	10.0
Skipping (2)	-0.055	.0245	.026	1.06	1.31
Proficiency (3)	0.038	.0148	.010	0.96	0.83

# Results

Correlations:

$$\begin{matrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \\ \widehat{\theta}_3 \end{matrix} \begin{pmatrix} 1.98 & & \\ 0.57 & 2.41 & \\ 0.53 & 0.56 & 0.86 \end{pmatrix}$$

$$\begin{matrix} \widehat{\varepsilon}_1 \\ \widehat{\varepsilon}_2 \\ \widehat{\varepsilon}_3 \end{matrix} \begin{pmatrix} 0.42 & & \\ 0.69 & 1.21 & \\ 0.29 & 0.31 & 0.74 \end{pmatrix}$$



# Results

- Clear effect of item position on *reaching*-threshold
- Missing data processes are related, but not the same as proficiency process.
  - MNAR
  - $r/d \neq 0$
- Threshold for reaching and skipping are related.

# Results

## Comparison with two other models:

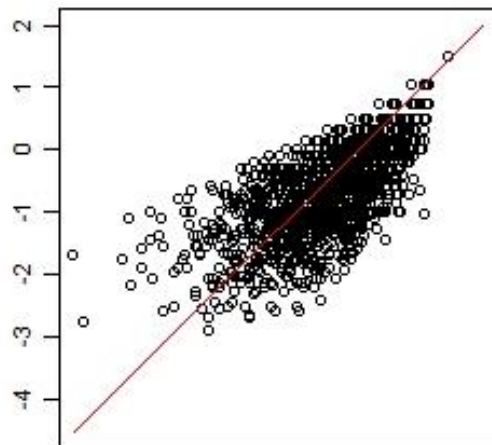
- Rasch Poisson
  - missing = wrong
  - counts model
  - number of items that are not correct
- Rasch model
  - Missing = ignorable
  - With item difficulties

# Results

Item Selection Model

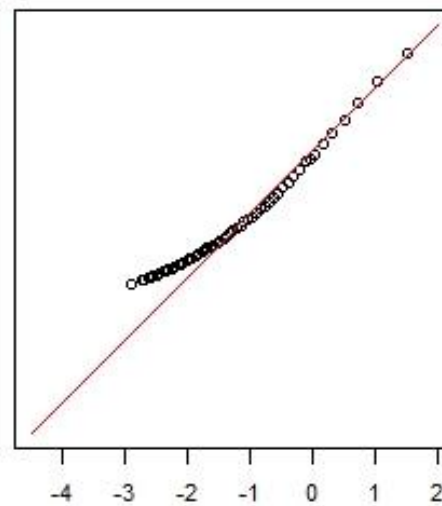
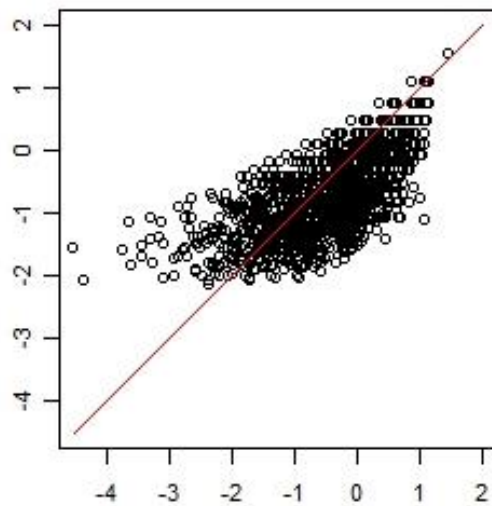
$$\rho = 0.82$$

$$\rho = 0.82$$



Rasch Model

$$\rho = 1$$



Poisson Model

# Content

1. Missing data
2. Handling missing data
3. IRTree
4. Two IRTrees for missing data
5. Illustration
6. Discussion

# Discussion

- Information in missing responses
- A lot of assumptions
- How much missing data is needed to get reliable estimates?
- Applicable in other domains?
- How to interpret the missingness processes?
  - Related to motivation,.. ?
  - Related to response times?
- When desirable?
  - Interested in missing data processes
  - Using information in missing responses
- Polytomous items / 2PL

# To remember

- Missing responses can be non-ignorable.
- But they can be modeled.
- We use IRTrees
  - ⇒ Can reduce possible bias
  - ⇒ Can help in understanding the processes underlying the missing data.

**Thank you for  
your attention.**

